PIVOT: Two-Phase Machine Translation System

Kazunori MURAKI
C&C Information Laboratories, NEC Corp.

NEC has developed and released a Japanese-English/English-Japanese machine translation system called PIVOT. This system translates from source to target language through the medium of an interlingua, using an analysis module and a synthesis module. This overview will describe some features of PIVOT, the concept of the interlingua used in PIVOT, and comment on it based on experience in the development of the system.

The system can be understood as an interaction between the analysis and synthesis processes and the structures in the lexicon. The analysis module maps input sentences to interlingua representations, using the source language dictionary in combination with grammar rules and semantic interpretation and knowledge rules. The synthesis module maps interlingua representations to surface sentences, using the target language dictionary and knowledge of target language grammar and style.

The symbols used in the interlingua are called CP's, or Conceptual Primitives. Except for symbols representing pragmatic and structural scope information, each CP is linked to various surface expressions (words or morphemes) in the dictionary for each language. Similarly, most surface expressions correspond to more than one CP. Thus, the main task of the analysis module is to choose the correct symbol to represent the meaning of each surface expression. Conversely, the task of the synthesis module is to choose a surface representation for each conceptual symbol.

The fundamental translation functions of the system described above are augmented for the user with the following features (ACOS System 4 environment features).

1: Batch translation features

A variety of functions are provided as tools.

Text processing and management
Lexicon development and maintenance
Dictionary and text printing

2: Interactive translation features

A variety of functions are provided through menus and interactive bilingual screens.

Bilingual text management
Dictionary updating and management

Bilingual editing
Automatic unknown-word detection, etc.

Basically, text input and editing is done using the text-processing facilities of a work station, while translation itself is done on the host mainframe computer.

It may be thought that the interactive translation environment is just a support system for hand translation, using the translation functions and the editing functions as tools. However, given the present functional immaturity of machine translation, many users find this to be a valuable way to work. Every user has a different way of using the system, and PIVOT's interactive environment is currently being improved in response to user suggestions.

The interlingual approach adopted in the construction of PIVOT has the advantage of being easily extendable to a multi-lingual system. In order to implement the interlingual approach, it is necessary to analyze and express semantic information which approaches such as transfer can ignore by using transfer rules and dictionary.

In addition to structural information from the input sentence and semantic information abstracted from the words and morphemes, PIVOT's interlingua represents pragmatic information such as topic and focus, and scope relations among semantic primitives.

This is summarized as follows. (1) through (4) represent abstract meaning derived from linguistic features. In (1), the arcs of the structural representation specify only dependency relations by the direction of the arc. This means the every information, even the relation like deep case in interlingua, can have conceptual meaning and can be paraphrased between structures and CP's by paraphrasing postulate(5).

1: Structural Information
   Acyclic Directional Network
   All nodes are associated with CP's.
   All arcs have a direction.
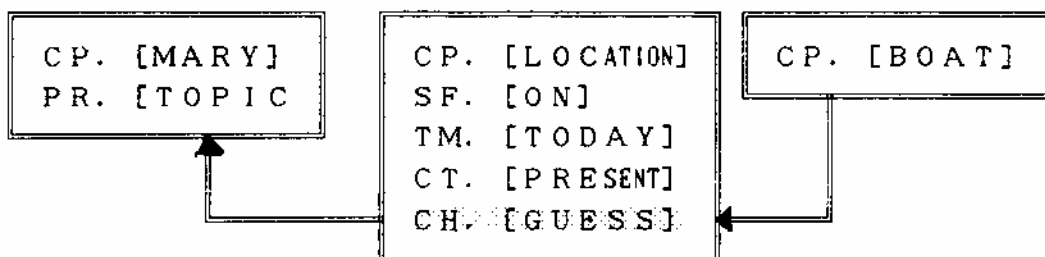
2: Node Information
> All nodes have several kinds of semantic information.
>
> Each piece of information is represented by a CP semantic symbol.
>> This information can include such concepts as (relative) location, (relative) time, aspect (active/stative), speaker/agent's intensionality, objects, things, relations, etc.

3: Pragmatic Information
> Topic, Focus
>
> Theme, Rheme
>
> Relative position with respect to the predicate

4: Scope Information
> Scope of comparison
>
> Scope of negation/quantification

5: CP-Structure Paraphrasing Postulates

These are coded by dictionary content and semantic postulates.

In building a translation system on this framework, important problems were encountered. The structural information (1) and node information (2) had to be designed in a language-independent way. Scope information (4) had to be represented in a format appropriate to the structure. A way of dealing with pragmatics (3) had to be found, though the treatment of such information has not yet been resolved in linguistics. Finally, semantic rules and dictionary representations had to be devised to allow paraphrase (5), for example to be able to translate between Japanese auxiliary verbs such as "rasii," which can be considered to have no (case) structure in Japanese, and the corresponding English verbs such as "seem," which do have case structure in English. This kind of paraphrasing should be done by monolingual dictionaries and meaning postulate without any kind of bilingual transfer dictionary and rules.

In the functional improvement of the PIVOT machine translation system, and further in the application of the interlingual approach to multilingual translation, the following three points are important.

1. More exhaustive and well defined set of semantic definitions of CPs
2. More sophisticated coding of pragmatics in the lexicon to strengthen selectivity both of CPs in analysis and surface words in generation
3. Logically better defined interaction between dictionary and rules

The first problem appeared much clearly when we tried to integrate the third language other than English and Japanese into our lexicon. Not well defined CPs made it too difficult or more impossible to set up links between the CPs and words in that language. This, we expect, will be resolved by energetic activities on semantics in EDR.

The second means the interpretive selectivity of CPs and words can be assured by extra linguistic and extra semantic information, what is called pragmatics in translation, as well as intra linguistic and semantic information. Especially, difference of each CP should be coded more explicitly. This, we expect, will be solved through much close collaboration with translation experts.

The last challenges us. We should continue the system improvement in every facet. In addition, this effort should be the one to inherit all kinds of user efforts on the current translation systems.